The Annals of Statistics
2004, Vol. 32, No. 2, 552–576
DOI: 10.1214/009053604000000094
© Institute of Mathematical Statistics, 2004

# MINIMAX ESTIMATION OF LINEAR FUNCTIONALS OVER NONCONVEX PARAMETER SPACES[1]

By T. Tony Cai and Mark G. Low

*University of Pennsylvania*

The minimax theory for estimating linear functionals is extended to the case of a finite union of convex parameter spaces. Upper and lower bounds for the minimax risk can still be described in terms of a modulus of continuity. However in contrast to the theory for convex parameter spaces rate optimal procedures are often required to be nonlinear. A construction of such nonlinear procedures is given. The results developed in this paper have important applications to the theory of adaptation.

**1. Introduction.** Let $Y$ be an observation from either the white noise model,

$$dY(t) = f(t)\,dt + n^{-1/2}\,dW(t) \tag{1}$$

where $W(t)$ is a standard Brownian motion, or the Gaussian sequence model

$$Y(i) = f(i) + n^{-1/2}\varepsilon_i \tag{2}$$

where $\varepsilon_i$ are i.i.d. standard normal random variables.

The minimax theory for estimating a linear functional $T$ has been studied in great generality when it is assumed that the function $f$ belongs to a parameter space which is convex. See, for example, Ibragimov and Has'minskii (1984), Donoho and Liu (1991a, b) and Donoho (1994). In particular, the properties of the minimax linear estimators can often be described precisely. In this case for any linear functional $T$ write $R_A^*(n;\mathcal{F})$ for the minimum (over all linear procedures) maximum mean squared error. Donoho and Liu (1991a) introduced a modulus of continuity

$$\omega(\varepsilon,\mathcal{F}) = \sup\{|T(g)-T(f)| : \|g-f\|_2 \le \varepsilon, f\in\mathcal{F}, g\in\mathcal{F}\}$$

Received February 2002; revised December 2002.

[1] Supported in part by NSF Grant DMS-03-06576.

*AMS 2000 subject classifications.* Primary 62G99; secondary 62F12, 62C20, 62M99.

*Key words and phrases.* Constrained risk inequality, linear functionals, minimax estimation, modulus of continuity, nonparametric functional estimation, white noise model.

where the norm in this equation is the $L_2$ norm in function space for the white noise with drift model and the $l_2$ norm in sequence space for the sequence model. Donoho and Liu (1991a, b) and Donoho (1994) have shown that in either of these two cases,

$$R_A^*(n;\mathcal{F}) = \sup_{\varepsilon>0} \omega^2(\varepsilon,\mathcal{F}) \frac{1/(4n)}{1/n+\varepsilon^2/4} \tag{3}$$

and that

$$\frac{1}{8}\omega^2\left(\frac{2}{\sqrt{n}},\mathcal{F}\right) \le R_A^*(n;\mathcal{F}) \le \frac{1}{4}\omega^2\left(\frac{2}{\sqrt{n}},\mathcal{F}\right).$$

An earlier version of this result can also be found in Ibragimov and Has'minskii (1984). Without the restriction to affine procedures write $R_N^*(n;\mathcal{F})$ for the minimax mean squared error for estimating the linear functional $T$. Donoho and Liu (1991b) have shown that

$$\frac{R_A^*(n;\mathcal{F})}{R_N^*(n;\mathcal{F})} \le 1.25.$$

Therefore the maximum risk of the optimal linear procedure is within a small constant factor of the minimax risk when the parameter space is convex. Of equal importance, Donoho and Liu (1991b) showed that the modulus can be used to give a recipe for constructing an affine procedure which has the maximum mean squared error attaining the bound given in (3).

Recent work on estimating linear functionals has focused on adaptive estimation. The goal is to find a single procedure which is near minimax simultaneously over a number of different parameter spaces. Pioneering work in this area began with Lepski (1990). This work focused on particularly important examples such as Lipschitz classes. In Efromovich and Low (1994) a general theory was developed for the case of nested convex parameter spaces.

A general extension of this adaptive estimation theory to spaces which are not nested must also include a minimax analysis for sets which are not convex. The reason for this is that we need to first know the minimax risk over the union of the original convex spaces and this space need not be convex unless the sets are nested. This paper focuses on such an extension of the minimax theory for estimating linear functionals over nonconvex parameter spaces. For applications to adaptive estimation see Cai and Low (2002). Although as just mentioned our primary motivation for this problem is the theory of adaptation the minimax theory itself is in fact quite interesting. In particular in this setting optimal linear procedures can sometimes have risks far from the optimal rate. In fact even if the parameter space is only a union of two convex sets it is possible that the maximum risk of the best linear estimator does not even converge even though the maximum risk of the optimal nonlinear procedure converges quickly. Such examples are given in Section 5.

Although optimal linear procedures need no longer be close to optimal we show that the minimax rate of convergence is still determined by the modulus of continuity over the parameter space when the parameter space is a finite union of convex sets. On the other hand, in Section 4, it is shown that the minimax linear risk is determined by the modulus of continuity over the convex hull of the parameter space. Therefore affine procedures fail when, in terms of the modulus, the convex hull is much larger than the parameter space itself. Such are the cases in the examples in Section 5. In these cases rate optimal estimators need to be nonlinear. A general construction of such nonlinear procedures is given in Section 3.

One of the main tools for the construction of the general procedure is a construction of linear procedures which have a given variance and precisely control the bias over two different convex parameter spaces. Upper bounds are given on the bias over one parameter space and lower bounds over the other. These linear procedures can then be used to test which of the convex sets the function lies in and then usual linear procedures can be used. The details of these arguments can be found in Sections 2 and 3.

The theoretical results are complemented by several illustrative examples given in Section 5 covering a range of cases. In the examples of estimating a linear functional of a nearly black object the parameter space is the union of a growing number of convex parameter spaces. In these cases the usual minimax lower bound is no longer sharp and the minimax rate of convergence is derived explicitly using a mixture prior and a constrained risk inequality.

## 2. Ordered modulus and bias variance tradeoffs.

One of the main tools for the construction of the general minimax procedure is the construction of linear procedures which have a given variance and precisely control the bias over two different convex parameter spaces. Upper bounds are given on the bias over one parameter space and lower bounds over the other. The key technical tool which allows for this construction is an ordered modulus of continuity between two function spaces. It is a generalization of the modulus of continuity introduced by Donoho and Liu (1991a) which has already been shown in Low (1995) to allow for the construction of a procedure which minimizes the maximum squared bias given a constraint on the maximum variance.

For a linear functional $T$ define an ordered modulus of continuity between two classes $\omega(\varepsilon, \mathcal{F}, \mathcal{G})$ by

$$\omega(\varepsilon, \mathcal{F}, \mathcal{G}) = \sup\{Tg - Tf : \|g - f\|_2 \leq \varepsilon; f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Note that $\omega(\varepsilon, \mathcal{F}, \mathcal{G})$ does not necessarily equal $\omega(\varepsilon, \mathcal{G}, \mathcal{F})$. It is clear that the modulus $\omega(\varepsilon, \mathcal{F}, \mathcal{G})$ is an increasing function of $\varepsilon$ and $0 \leq \omega(\varepsilon, \mathcal{F}, \mathcal{G}) \leq \infty$

if $\mathcal{F} \cap \mathcal{G} \neq \varnothing$. The between class modulus is also instrumental in the analysis of adaptation over different parameter spaces [see Cai and Low (2002)].

When $\mathcal{G} = \mathcal{F}$, $\omega(\varepsilon, \mathcal{F}, \mathcal{F})$ is the usual modulus of continuity over $\mathcal{F}$ and will be denoted by $\omega(\varepsilon, \mathcal{F})$. The following result on the concavity of the modulus is important in the bias variance tradeoffs and in the construction of the general minimax procedure.

THEOREM 1. *Assume that $\mathcal{F}$, $\mathcal{G}$ are convex and that $\mathcal{F} \cap \mathcal{G} \neq \varnothing$. Let $T$ be a linear functional. Then the function $\omega(\varepsilon, \mathcal{F}, \mathcal{G})$ is a concave function of $\varepsilon$. In particular it follows that, for $D > 1$,*

$$\omega(D\varepsilon, \mathcal{F}, \mathcal{G}) \leq D\omega(\varepsilon, \mathcal{F}, \mathcal{G}).$$

PROOF. Suppose that $g_1 \in \mathcal{G}$, $g_2 \in \mathcal{G}$ and $f_1 \in \mathcal{F}$, $f_2 \in \mathcal{F}$ with

$$\|g_i - f_i\|_2 \leq \varepsilon_i.$$

Then, for $0 \leq \lambda \leq 1$,

$$\|\lambda g_2 + (1-\lambda) g_1 - [\lambda f_2 + (1-\lambda) f_1]\|_2 \leq \lambda \varepsilon_2 + (1-\lambda)\varepsilon_1$$

and

$$\begin{aligned} &T(\lambda g_2 + (1-\lambda) g_1 - [\lambda f_2 + (1-\lambda) f_1]) \\ &\qquad = \lambda(T(g_2) - T(f_2)) + (1-\lambda)(T(g_1) - T(f_1)). \end{aligned}$$

It then follows that

$$\omega(\lambda\varepsilon_2 + (1-\lambda)\varepsilon_1, \mathcal{F}, \mathcal{G}) \geq \lambda\omega(\varepsilon_2, \mathcal{F}, \mathcal{G}) + (1-\lambda)\omega(\varepsilon_1, \mathcal{F}, \mathcal{G})$$

and so $\omega$ is concave. □

As mentioned earlier in Low (1995) it was shown that in the white noise model for any linear functional the modulus of continuity can be used to precisely trade off various levels of bias and variance over a given convex parameter space. The modulus of continuity between parameter spaces can be used to perform an analogous trade. It can be used to give a linear procedure which has upper bounds for the bias over one parameter space and lower bounds for the bias over the other parameter space. The detailed results are given in Theorems 2 and 3 below.

We shall write $\langle u, v\rangle$ for the usual $l_2$ inner product for either sequence or function space. Specifically if we observe the white noise with drift model let

$$\langle f, g\rangle = \int fg$$

and if we observe the sequence model let

$$\langle f, g\rangle = \sum f_i g_i.$$

For all $V \geq 0$ let

$$B(V, \mathcal{F}, \mathcal{G}) = 2^{-1} \sup_{\varepsilon > 0} (\omega(\varepsilon, \mathcal{F}, \mathcal{G}) - \sqrt{nV}\,\varepsilon). \tag{4}$$

It will also be convenient to introduce an inverse of $B(V, \mathcal{F}, \mathcal{G})$ defined for all $B \geq 0$ by

$$V(B, \mathcal{F}, \mathcal{G}) = \sup_{\varepsilon > 0} \frac{1}{n\varepsilon^2} ([\omega(\varepsilon, \mathcal{F}, \mathcal{G}) - 2B]_+)^2. \tag{5}$$

We shall show in Theorems 2 and 3 that there is a linear estimator with variance bounded by $V$, which has maximum bias over $\mathcal{F}$ less than or equal to $B(V, \mathcal{F}, \mathcal{G})$ and minimum bias over $\mathcal{G}$ greater than or equal to $-B(V, \mathcal{F}, \mathcal{G})$. Theorem 2 covers the most usual situations where linear estimators can be easily described in terms of the modulus. Theorem 3 extends the theory to cover the general case.

Our analysis is split into a number of cases. The most usual ones are covered by cases 1(a) and 2(a). It is these cases which are in fact needed in the construction of the general procedure in Section 3. We include the others for completeness. First note that we shall always assume that $\omega(1, \mathcal{F}, \mathcal{G}) > 0$; otherwise the linear functional is constant over $\mathcal{F} \cup \mathcal{G}$ and the estimation problem is thus trivial.

Case 1. Suppose that $0 < B(V, \mathcal{F}, \mathcal{G}) < \infty$. Then define $\varepsilon(V, \mathcal{F}, \mathcal{G})$ by

$$\varepsilon(V, \mathcal{F}, \mathcal{G}) = \arg\max_{\varepsilon \geq 0} (\omega(\varepsilon, \mathcal{F}, \mathcal{G}) - \sqrt{nV}\varepsilon) \tag{6}$$

where $\varepsilon(V, \mathcal{F}, \mathcal{G})$ is the smallest value of $\varepsilon$ for which the maximum in (4) is attained. It will be convenient to break case 1 into two further cases, namely:

(a) $0 < \varepsilon(V, \mathcal{F}, \mathcal{G}) < \infty$.
(b) $\varepsilon(V, \mathcal{F}, \mathcal{G}) = \infty$.

Case 2. $B(V, \mathcal{F}, \mathcal{G}) = 0$ and $B(V', \mathcal{F}, \mathcal{G}) > 0$ for all $0 \leq V' < V$. Note that if $B(V', \mathcal{F}, \mathcal{G}) = 0$ for some $V' < V$ then we could reduce the variance of our estimator without increasing the magnitude of the bias. Under this assumption there are only two possibilities.

(a) $\omega(\varepsilon, \mathcal{F}, \mathcal{G}) = \sqrt{nV}\varepsilon$ on some interval $0 \leq \varepsilon \leq \varepsilon_0$ where $\varepsilon_0 > 0$. We can then define $\varepsilon(V, \mathcal{F}, \mathcal{G})$ to be the largest $\varepsilon \leq \frac{1}{\sqrt{n}}$ for which $\omega(\varepsilon, \mathcal{F}, \mathcal{G}) = \sqrt{nV}\varepsilon$.

(b) $\omega(\varepsilon, \mathcal{F}, \mathcal{G}) < \sqrt{nV}\varepsilon$ whenever $\varepsilon > 0$. It then follows from the concavity of the modulus that $0 < B(V', \mathcal{F}, \mathcal{G}) < \infty$ for some $V' < V$. In this case set $\varepsilon(V, \mathcal{F}, \mathcal{G}) = 0$.

The following technical lemma shows that $B(V, \mathcal{F}, \mathcal{G})$ is continuous in $V$ whenever it is finite.

LEMMA 1. *Suppose $\mathcal{F}$ and $\mathcal{G}$ are closed and convex with $\mathcal{F} \cap \mathcal{G} \neq \varnothing$. Then $\varepsilon(V, \mathcal{F}, \mathcal{G})$ is nonincreasing in $V$. Assume $B(V, \mathcal{F}, \mathcal{G}) < \infty$. Then $\varepsilon(V', \mathcal{F}, \mathcal{G}) < \infty$ if $V' > V$ and*

$$\lim_{V_m \downarrow V} B(V_m, \mathcal{F}, \mathcal{G}) = B(V, \mathcal{F}, \mathcal{G}). \tag{7}$$

*If, in addition, $B(V', \mathcal{F}, \mathcal{G}) < \infty$ for some $V' < V$, then*

$$\lim_{V_m \to V} B(V_m, \mathcal{F}, \mathcal{G}) = B(V, \mathcal{F}, \mathcal{G}). \tag{8}$$

PROOF. Note first that the monotonicity of $\varepsilon(V, \mathcal{F}, \mathcal{G})$ and the fact that $\varepsilon(V', \mathcal{F}, \mathcal{G}) < \infty$ if $V' > V$ follows from the concavity of the modulus $\omega(\varepsilon, \mathcal{F}, \mathcal{G})$ as shown in Theorem 1. Now assume that $B(V, \mathcal{F}, \mathcal{G}) < \infty$ and let $V_m \downarrow V$. Note that

$$B(V_m, \mathcal{F}, \mathcal{G}) \le B(V, \mathcal{F}, \mathcal{G}) \qquad \text{for } V_m \ge V,$$

and that for any $\varepsilon$

$$B(V_m, \mathcal{F}, \mathcal{G}) \ge 2^{-1}(\omega(\varepsilon, \mathcal{F}, \mathcal{G}) - \sqrt{nV_m}\varepsilon).$$

Taking limits yields

$$\liminf_{V_m \downarrow V} B(V_m, \mathcal{F}, \mathcal{G}) \ge 2^{-1}(\omega(\varepsilon, \mathcal{F}, \mathcal{G}) - \sqrt{nV}\varepsilon)$$

for all $\varepsilon$ and so taking the supremum over all $\varepsilon$ on the right-hand side shows that the limit exists and is equal to $B(V, \mathcal{F}, \mathcal{G})$. This proves (7).

Note that $B(V, \mathcal{F}, \mathcal{G})$ is a convex function of $\sqrt{V}$ since it is a supremum of a collection of convex functions of $\sqrt{V}$. Hence $B(V, \mathcal{F}, \mathcal{G})$ is continuous in $V$ on any open interval over which it is finite. Hence if $B(V', \mathcal{F}, \mathcal{G}) < \infty$ for some $V' < V$ then $B(\cdot, \mathcal{F}, \mathcal{G})$ is continuous at $V$ and so (8) follows. □

We now state the bias—variance tradeoff theorem in the most easily understood and most typical case where $0 < \varepsilon(V, \mathcal{F}, \mathcal{G}) < \infty$ and the modulus is attained by two functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$.

THEOREM 2. *Suppose $\mathcal{F}$ and $\mathcal{G}$ are convex and closed with $\mathcal{F} \cap \mathcal{G} \neq \varnothing$. Assume that $0 < \varepsilon(V, \mathcal{F}, \mathcal{G}) < \infty$. Suppose further that there are $f \in \mathcal{F}, g \in \mathcal{G}$ such that*

$$\|g - f\|_2 = \varepsilon(V, \mathcal{F}, \mathcal{G}) \equiv \varepsilon_V \quad \textit{and} \quad Tg - Tf = \omega(\varepsilon_V, \mathcal{F}, \mathcal{G}). \tag{9}$$

*Write* $u \equiv \frac{g-f}{\varepsilon_V}$ *for the direction of the affine family joining* $g$ *and* $f$. *Let*

$$a = T\left(\frac{f+g}{2}\right) - \sqrt{nV}\left\langle u, \frac{f+g}{2}\right\rangle. \tag{10}$$

*Then the estimator*

$$\hat{T}_V = a + \sqrt{nV}\int u(t)\,dY(t) \tag{11}$$

*for the white noise with drift model and the estimator*

$$\hat{T}_V = a + \sqrt{nV}\sum u(i)Y(i) \tag{12}$$

*for the sequence model have constant variance*

$$E(\hat{T}_V - E\hat{T}_V)^2 = V \tag{13}$$

*and have biases bounded by*

$$\sup_{f\in\mathcal{F}} E_f\hat{T}_V - Tf = B(V,\mathcal{F},\mathcal{G}) \tag{14}$$

*and*

$$\inf_{g\in\mathcal{G}} E_g\hat{T}_V - Tg = -B(V,\mathcal{F},\mathcal{G}). \tag{15}$$

Remark. If $\mathcal{F}$ and $\mathcal{G}$ are closed, convex and norm bounded with nonempty intersection then the condition that the modulus is attained is guaranteed. The extension to cases where either the modulus is not attained as well as for when $\varepsilon(V,\mathcal{F},\mathcal{G}) = 0$ and $\varepsilon(V,\mathcal{F},\mathcal{G}) = \infty$ will be covered in Theorem 3.

Proof of Theorem 2. The proof of this theorem essentially follows that of Theorem 2 in Low (1995). Note that the proofs of (14) and (15) are entirely similar so we shall only give the details for the proof of (15).

Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be extremal functions satisfying (9) which exist since $\mathcal{F}$ and $\mathcal{G}$ are closed. Let $h$ be any other element of $\mathcal{G}$. The affine family joining $g$ and $h$ is given by $(1-\theta)g + \theta h$, $0 \le \theta \le 1$. Let

$$J(\theta) = T((1-\theta)g + \theta h) - Tf - \sqrt{nV}\|(1-\theta)g + \theta h - f\|_2.$$

It follows from the definition of $\varepsilon(V,\mathcal{F},\mathcal{G})$ given in (6) that $J(\theta) \le J(0)$ for all $0 \le \theta \le 1$ and since $J(\theta)$ is clearly differentiable it follows that $J'(0) \le 0$. A simple computation shows that

$$Th - Tg - \sqrt{nV}\langle u, (h-g)\rangle \le 0. \tag{16}$$

Now

$$E\hat{T}_V - Tg = T\left(\frac{f+g}{2}\right) + \sqrt{nV}\left\langle u, \left(g - \frac{f+g}{2}\right)\right\rangle - Tg \tag{17}$$

and

$$E\hat{T}_V - Th = T\left(\frac{f+g}{2}\right) + \sqrt{nV}\left\langle u, \left(h - \frac{f+g}{2}\right)\right\rangle - Th. \tag{18}$$

It then follows from (16)–(18) that

$$(E\hat{T}_V - Tg) - (E\hat{T}_V - Th) \le 0. \tag{19}$$

Finally note that a simple calculation yields

$$E\hat{T}_V - Tg = -B(V, \mathcal{F}, \mathcal{G}). \tag{20}$$

Equations (19) and (20) combine to show (15) and the proof is complete. □

Theorem 2 treats the cases 1(a) and 2(a) under the additional assumption that the modulus is attained by $f \in \mathcal{F}$ and $g \in \mathcal{G}$. The functions $f$ and $g$ are used explicitly in the construction of the estimate $\hat{T}_V$. In general, the modulus may not be attained and in these cases the description of a linear estimator which trades variance and bias is more involved. We describe the general case in detail in the following theorem. Some of the details are similar to those given in Section 12 of Donoho (1994).

Define $B(m)$ to be the closed $L_2$ ball with radius $m$ and let $\mathcal{F}_m = \mathcal{F} \cap B(m)$ and $\mathcal{G}_m = \mathcal{G} \cap B(m)$. It follows from Lemma 2 of Donoho (1994) that for $\mathcal{F}_m$ and $\mathcal{G}_m$ the modulus $\omega(\varepsilon, \mathcal{F}_m, \mathcal{G}_m)$ can always be attained by some $f \in \mathcal{F}_m$ and $g \in \mathcal{G}_m$.

Define $V_m$, $\varepsilon_m$, $f_m$ and $g_m$ in the following way.

Case 1.

(a) $0 < B(V, \mathcal{F}, \mathcal{G}) < \infty$ and $0 < \varepsilon(V, \mathcal{F}, \mathcal{G}) < \infty$. In this case let $V_m = V$, $l(m) = m$ and define $\varepsilon_m = \varepsilon(V_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)})$. Note that for large $m$, $\varepsilon_m > 0$. Moreover, since both $\mathcal{F}_m$ and $\mathcal{G}_m$ are contained in $B(m)$ it follows that $\varepsilon_m < 2m$. Since $\mathcal{F}_{l(m)}$ and $\mathcal{G}_{l(m)}$ are closed and norm bounded it follows from Lemma 2 of Donoho (1994) that the modulus $\omega(\varepsilon_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)})$ is attained by a pair $f_m \in \mathcal{F}_{l(m)}$ and $g_m \in \mathcal{G}_{l(m)}$.

(b) $\varepsilon(V, \mathcal{F}, \mathcal{G}) = \infty$. In this case let $V_m > V$ be chosen where $V_m \downarrow V$. Then it follows from Lemma 1 that $B(V_m, \mathcal{F}, \mathcal{G}) \to B(V, \mathcal{F}, \mathcal{G})$. So for large $m$, $0 < B(V_m, \mathcal{F}, \mathcal{G}) < \infty$. Now choose an increasing sequence $l(m) \to \infty$ so that $B(V_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)}) > 0$. Now define $\varepsilon_m = \varepsilon(V_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)})$ and once again note that for large $m$, $0 < \varepsilon_m < 2m$. Again $\mathcal{F}_{l(m)}$ and $\mathcal{G}_{l(m)}$ are closed and norm bounded so the modulus $\omega(\varepsilon_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)})$ is attained by a pair $f_m \in \mathcal{F}_{l(m)}$ and $g_m \in \mathcal{G}_{l(m)}$.

Case 2.

(a) $B(V,\mathcal{F},\mathcal{G})=0$, $B(V',\mathcal{F},\mathcal{G})>0$ for all $V'<V$ and $\omega(\varepsilon,\mathcal{F},\mathcal{G})=\sqrt{nV}\varepsilon$ on some interval $0\le\varepsilon\le\varepsilon_0$ for some $\varepsilon_0>0$. Let $l(m)=m$ and note that at least for $m$ sufficiently large $0<B(0,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})<\infty$ and that since $\mathcal{F}_{l(m)}\subseteq\mathcal{F}$ and $\mathcal{G}_{l(m)}\subseteq\mathcal{G}$ it also follows that $B(V,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})=0$. Lemma 1 shows that for all sufficiently large $m$ there exists a $V_m<V$ such that

$$0<B(V_m,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})\le\frac{1}{m}.$$

Now let $\varepsilon_m=\varepsilon(V_m,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})$. Then as before it follows that for large $m$, $0<\varepsilon_m<2m$. Now since $\mathcal{F}_{l(m)}$ and $\mathcal{G}_{l(m)}$ are closed and norm bounded, the modulus $\omega(\varepsilon_m,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})$ is attained by a pair $f_m\in\mathcal{F}_{l(m)}$ and $g_m\in\mathcal{G}_{l(m)}$.

(b) $B(V,\mathcal{F},\mathcal{G})=0$, $B(V',\mathcal{F},\mathcal{G})>0$ for all $0\le V'<V$, $\omega(\varepsilon,\mathcal{F},\mathcal{G})<\sqrt{nV}\varepsilon$ whenever $\varepsilon>0$.

Now let $V_m<V$ be chosen where $V_m\uparrow V$. Note that there exists some $V_0\ge0$ such that

$$0<B(V',\mathcal{F},\mathcal{G})<\infty$$

for $V_0\le V'<V$. Then for large $m$,

$$0<B(V_m,\mathcal{F},\mathcal{G})<\infty.$$

So there is an increasing sequence $l(m)\to\infty$ such that

$$0<B(V_m,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})<B(V_m,\mathcal{F},\mathcal{G})<\infty.$$

We now define $\varepsilon_m=\varepsilon(V_m,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})$. It follows once again that $0<\varepsilon_m<2m$ for large $m$. Now since $\mathcal{F}_{l(m)}$ and $\mathcal{G}_{l(m)}$ are closed and norm bounded the modulus $\omega(\varepsilon_m,\mathcal{F}_{l(m)},\mathcal{G}_{l(m)})$ is attained by a pair $f_m\in\mathcal{F}_{l(m)}$ and $g_m\in\mathcal{G}_{l(m)}$.

For $V_m$, $\varepsilon_m$, $f_m$ and $g_m$ as just defined let $u_m=\frac{g_m-f_m}{\varepsilon_m}$ and let

$$a_m=T\left(\frac{f_m+g_m}{2}\right)-\sqrt{nV_m}\left\langle u_m,\frac{f_m+g_m}{2}\right\rangle.$$

For the white noise with drift model let

$$\hat{T}_m=a_m+\sqrt{nV_m}\int u_m(t)\,dY(t) \tag{21}$$

and for the sequence model let

$$\hat{T}_m=a_m+\sqrt{nV_m}\sum_i u_m(i)Y(i). \tag{22}$$

The estimator $\hat{T}_m$ corresponds to the estimator $\hat{T}_V$ defined in Theorem 2 for $V=V_m$, $\mathcal{F}=\mathcal{F}_{l(m)}$ and $\mathcal{G}=\mathcal{G}_{l(m)}$. In the general case we need to take a limit of the estimators $\hat{T}_m$.

Note that $\|u_m\|_2 = 1$ and so there exists a subsequence which converges weakly to some function $u$ where $\|u\|_2 \leq 1$.

Now let $h \in \mathcal{F} \cap \mathcal{G}$. Then since $\|h\|_2 \leq m_0$ for some $m_0 < \infty$ it follows that $h \in \mathcal{F}_{l(m)} \cap \mathcal{G}_{l(m)}$ for all $m \geq m_1$ where $l(m_1) \geq m_0$.

For $m \geq m_1$ note it follows from Theorem 2 that

$$|E_h \hat{T}_m - Th| \leq B(V_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)}) \leq \frac{1}{m}$$

in case 2(a), and in all other cases

$$|E_h \hat{T}_m - Th| \leq B(V_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)}) \leq B(V_m, \mathcal{F}, \mathcal{G}).$$

Note that $B(V_m, \mathcal{F}, \mathcal{G})$ is bounded since it converges to $B(V, \mathcal{F}, \mathcal{G})$. Note also that

$$E\hat{T}_m = a_m + \sqrt{nV_m}\langle u_m, h\rangle$$

and since the norm of $u_m$ is equal to one it follows that $a_m$ is bounded. Hence there is a subsequence of the subsequence used to define $u$ which converges to some finite $a$. Denote this subsubsequence by $m_k^*$.

For the white noise with drift model let

$$\hat{T}_V = a + \sqrt{nV} \int u(t)\, dY(t) \tag{23}$$

and for the sequence model let

$$\hat{T}_V = a + \sqrt{nV} \sum u(i) Y(i). \tag{24}$$

The following theorem shows that this estimator $\hat{T}_V$ which has been formed as a limit of $\hat{T}_m$ trades bias and variance in the general case.

Theorem 3. *Suppose $\mathcal{F}$ and $\mathcal{G}$ are convex and closed with nonempty intersection. Then the estimator defined by (23) for the white noise with drift model and (24) for the sequence model satisfies*

$$E(\hat{T}_V - E\hat{T}_V)^2 \leq V \tag{25}$$

*and has biases bounded by*

$$\sup_{f \in \mathcal{F}} E_f \hat{T}_V - Tf \leq B(V, \mathcal{F}, \mathcal{G}) \tag{26}$$

*and*

$$\inf_{g \in \mathcal{G}} E_g \hat{T}_V - Tg \geq -B(V, \mathcal{F}, \mathcal{G}). \tag{27}$$

PROOF. Note that (25) follows immediately from the fact that the norm of $u$ is bounded by 1. We shall only give the proof for (26) since the proofs for the other cases are analogous.

First note that the estimator $\hat{T}_m$ as defined in (21) and (22) satisfies the bounds given in Theorem 2. If $f \in \mathcal{F}_{l(m)}$ then

$$E\hat{T}_m - Tf \le B(V_m, \mathcal{F}_{l(m)}, \mathcal{G}_{l(m)}).$$

Let $m_k^*$ be the subsubsequence along which $a_m$ and $u_m$ converge to $a$ and $u$, respectively. Now for any $f \in \mathcal{F}$, $f \in \mathcal{F}_{l(m_k^*)}$ for large $k$. So

$$\begin{aligned} E\hat{T}_V - Tf &\le \limsup_{k\to\infty} E\hat{T}_{m_k^*} - Tf \\ &\le \limsup_{k\to\infty} B(V_{m_k^*}, \mathcal{F}_{l(m_k^*)}, \mathcal{G}_{l(m_k^*)}) \\ &\le \limsup_{k\to\infty} B(V_{m_k^*}, \mathcal{F}, \mathcal{G}) \\ &\le B(V, \mathcal{F}, \mathcal{G}). \end{aligned}$$

The last step follows from Lemma 1. □

REMARK. Using the Cramér–Rao inequality arguments found in Low (1995) it can be shown that the linear estimator which attains the bounds in the theorem is in fact unique and must actually attain the inequalities. It then follows that the sequence $u_m$ which was used to define the estimator $\hat{T}_V$ actually converges strongly to $u$ and that the sequence $a_m$ actually converges.

## 3. Minimax estimator over a finite union of convex sets.

Let $\mathcal{F} = \bigcup_{i=1}^k \mathcal{F}_i$ where for $i = 1, \dots, k$, $\mathcal{F}_i$ are closed convex spaces with nonempty intersections, that is, $\mathcal{F}_i \cap \mathcal{F}_j \neq \varnothing$ for all $i, j$. Our objective is to construct an estimator which is rate optimal for estimating a linear functional $T$ over the parameter space $\mathcal{F}$. Standard two-point testing arguments as, for example, contained in Donoho and Liu (1991a) or Brown and Low (1996) show that the minimax risk for estimating a linear functional $Tf$ over $\mathcal{F}$ is bounded from below by

$$\inf_{\hat{T}} \sup_{f\in\mathcal{F}} E(\hat{T} - Tf)^2 \ge \frac{1}{8}\omega^2\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right). \tag{28}$$

Let $\hat{T}_i$ be linear estimators which satisfy

$$\sup_{f\in\mathcal{F}_i} E(\hat{T}_i - Tf)^2 \le M^2\omega^2\left(\frac{1}{\sqrt{n}}, \mathcal{F}_i\right) \tag{29}$$

for some $M > 0$. As mentioned in the introduction if $M \ge 1$ such linear estimators are guaranteed to exist and can be constructed by the recipe given in Donoho (1994). In the following discussion $C$ will denote generic constants whereas $M$ will always refer to the bounds given in (29).

For $i \neq j$, let $V_{i,j} = \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)$. Then it follows from the concavity of the modulus that $B(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j)$ as defined by (4) satisfies

$$\begin{aligned} B(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) &= 2^{-1} \sup_{\varepsilon > 0} (\omega(\varepsilon, \mathcal{F}_i, \mathcal{F}_j) - \sqrt{n V_{i,j}} \varepsilon) \\ &= 2^{-1} \sup_{\varepsilon \le 1/\sqrt{n}} \left( \omega(\varepsilon, \mathcal{F}_i, \mathcal{F}_j) - \sqrt{n} \varepsilon \omega\left( \frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j \right) \right) \\ &\le 2^{-1} \omega\left( \frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j \right). \end{aligned}$$

Hence either $B(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) = 0$ or $0 < B(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) \le 2^{-1} \omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)$. In the first case when $B(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) = 0$ it follows from the definition of $\varepsilon(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j)$ given for case 2(a) that $\varepsilon(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) = \frac{1}{\sqrt{n}}$. On the other hand if $0 < B(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) \le 2^{-1} \omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)$ then $0 < \varepsilon(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) \le \frac{1}{\sqrt{n}}$. Hence we know in both cases that $0 < \varepsilon(V_{i,j}, \mathcal{F}_i, \mathcal{F}_j) \le \frac{1}{\sqrt{n}}$. It follows that when using $V_{i,j} = \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)$ that we are in either case 1(a) or case 2(a) of Section 2.

For $i \neq j$ let $\hat{T}_{i,j}$ be the estimator defined as in Theorem 2 when (9) is attained where $\mathcal{F} = \mathcal{F}_i$, $\mathcal{G} = \mathcal{F}_j$ and $V = V_{i,j} = \omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)$. When (9) is not attained the estimator $\hat{T}_{i,j}$ is defined as in Theorem 3. This linear estimator has variance bounded by $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)$ and bias which satisfies

$$-2^{-1} \omega\left( \frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j \right) \le \inf_{f \in \mathcal{F}_j} (E(\hat{T}_{i,j}) - Tf) \tag{30}$$

and

$$\sup_{f \in \mathcal{F}_i} (E(\hat{T}_{i,j}) - Tf) \le 2^{-1} \omega\left( \frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j \right). \tag{31}$$

Now based on the linear estimators $\hat{T}_{i,j}$ and the linear estimators $\hat{T}_i$, which satisfy (29), define $\hat{z}^u_{i,j}$, $\hat{z}^l_{i,j}$ and $\hat{z}_{i,j}$ by

$$\hat{z}^u_{i,j} = \frac{\hat{T}_{i,j} - \hat{T}_i}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j) + M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i)},$$

$$\hat{z}^l_{i,j} = \frac{\hat{T}_i - \hat{T}_{j,i}}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_j, \mathcal{F}_i) + M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i)}$$

and

$$\hat{z}_{i,j} = \max(\hat{z}^u_{i,j}, \hat{z}^l_{i,j}).$$

Note that $\hat{z}^u_{i,j}$ and $\hat{z}^l_{i,j}$ are normally distributed and satisfy

$$\max(\operatorname{Var}(\hat{z}^u_{i,j}), \operatorname{Var}(\hat{z}^l_{i,j})) \le 1. \tag{32}$$

Finally define the estimator of the linear functional $T$ as

$$\hat{T}^* = \hat{T}_{\hat{i}} \qquad \text{with } \hat{i} = \arg\min_i \Bigl(\sup_{j \ne i} \hat{z}_{i,j}\Bigr). \tag{33}$$

The analysis of the mean squared error of $\hat{T}^*$ is facilitated by the following lemma which bounds the probability that $\hat{T}^* = \hat{T}_j$ when the magnitude of the bias of $\hat{T}_j$ is large.

Lemma 2. *Suppose $f \in \mathcal{F}_i$ and for some $j \ne i$, $|E\hat{T}_j - Tf| \ge \gamma M \omega(\frac{1}{\sqrt{n}}, \mathcal{F})$ where $\gamma \ge 3$. Then*

$$P(\hat{i} = j) \le 2k \exp\biggl(-\frac{(\gamma-3)^2}{32}\biggr).$$

Proof. Note that if $f \in \mathcal{F}_i$ and $j \ne i$, then from (29) and (31),

$$E\hat{z}^u_{i,j} = \frac{E(\hat{T}_{i,j} - Tf) - E(\hat{T}_i - Tf)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j) + M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i)} \le 1 \tag{34}$$

and from (29) and (30),

$$E\hat{z}^l_{i,j} = \frac{E(\hat{T}_i - Tf) - E(\hat{T}_{j,i} - Tf)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_j, \mathcal{F}_i) + M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i)} \le 1. \tag{35}$$

Now suppose that $f \in \mathcal{F}_i$. We shall only give details of the proof when

$$E\hat{T}_j - Tf \ge \gamma M \omega\biggl(\frac{1}{\sqrt{n}}, \mathcal{F}\biggr)$$

as the case when

$$E\hat{T}_j - Tf \le -\gamma M \omega\biggl(\frac{1}{\sqrt{n}}, \mathcal{F}\biggr)$$

is handled in a similar way. When $E\hat{T}_j - Tf \ge \gamma M\omega(\frac{1}{\sqrt{n}}, \mathcal{F})$ then it follows from (31) that

$$\begin{aligned} E\hat{z}^l_{j,i} &= \frac{E(\hat{T}_j - Tf) - E(\hat{T}_{i,j} - Tf)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j) + M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_j)} \\ &\ge \frac{\gamma M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}) - \frac{1}{2}\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j)}{\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_i, \mathcal{F}_j) + M\omega(\frac{1}{\sqrt{n}}, \mathcal{F}_j)} \\ &\ge \frac{\gamma - 1}{2}. \end{aligned} \tag{36}$$

Now without loss of generality suppose that $i = 1$ and that $j = 2$. Then if $\hat{i} = 2$ note that $\hat{z}_{2,1} \le \sup_{r \ne 1}(\hat{z}_{1,r})$ and since $\hat{z}^l_{2,1} \le \hat{z}_{2,1}$ it follows that

$$\begin{aligned} P(\hat{i} = 2) &\le \sum_{r=2}^{k} P(\hat{z}^l_{2,1} - \hat{z}_{1,r} \le 0) \\ &\le \sum_{r=2}^{k} \{P(\hat{z}^l_{2,1} - \hat{z}^l_{1,r} \le 0) + P(\hat{z}^l_{2,1} - \hat{z}^u_{1,r} \le 0)\}. \end{aligned}$$

Now by (32) $\hat{z}^l_{2,1} - \hat{z}^l_{1,r}$ and $\hat{z}^l_{2,1} - \hat{z}^u_{1,r}$ both have normal distributions with variance less than or equal to 4 and by (34)–(36) means greater than or equal to $\frac{\gamma-3}{2}$ and the lemma now follows from the bound on a standard normal random variable $Z$,

$$P(Z > t) \le \exp\left(-\frac{t^2}{2}\right)$$

which holds for all $t \ge 0$. □

Although our main focus is on mean squared error we shall consider the more general case of $p$th power loss. Such general cases are important in the theory of adaptation [see Cai and Low (2002)]. Lemma 2 can be used to bound the risk of the estimator $\hat{T}^*$ defined by (33) as in the following theorem.

THEOREM 4. *Suppose either the white noise model* (1) *or the sequence model* (2) *is given. Let* $\mathcal{F} = \bigcup_{i=1}^k \mathcal{F}_i$ *where* $k \ge 2$ *and* $\mathcal{F}_i$ *are closed convex sets with* $\mathcal{F}_i \cap \mathcal{F}_j \ne \varnothing$ *for all* $i, j$. *Let* $\hat{T}^*$ *be the estimator of the linear functional* $T$ *defined as in* (33). *Then for* $p \ge 1$,

$$\sup_{f \in \mathcal{F}} E|\hat{T}^* - Tf|^p \le C(p) M^p (\ln k)^{p/2} \omega^p\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right) \tag{37}$$

*where the constant* $C(p)$ *is independent of* $M$, $k$ *and* $n$.

REMARK. Note that we can always find linear estimators $\hat{T}_i$ for which $M \le 1$ in (29) and so the theorem yields an upper bound on the minimax risk over $\mathcal{F}$ which only depends on the modulus and the number $k$. There is also a minimax lower bound for the $p$th power loss analogous to that given in (28),

$$\inf_{\hat{T}} \sup_{f \in \mathcal{F}} E|\hat{T} - Tf|^p \ge b(p) \omega^p\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right) \tag{38}$$

for some constant $b(p) > 0$. By comparing the upper bound in (37) with this bound it is clear that for fixed finite $k$ the estimator $\hat{T}^*$ is rate optimal over

$\mathcal{F}$. It is also worth noting that sometimes the lower bound is not asymptotically sharp when $k$ is finite but grows with $n$. In Section 5 we give examples where $k$ grows with $n$ and the optimal rate is given by the upper bound in equation (37).

In the theory of adaptation the goal is to find a procedure which is simultaneously near minimax over a collection of parameter spaces. If a collection of convex parameter spaces is not nested then the largest of the minimax risks for each convex parameter space may be smaller than the minimax risk over the union of the convex parameter spaces [see, e.g., Cai and Low (2002)]. In such cases an appropriate benchmark for the maximum risk of an adaptive estimator is given by the bound in Theorem 4.

The proof of Theorem 4 is based on Lemma 2 and the following bound on the tail probabilities of a maximum of Gaussian random variables.

LEMMA 3. *Let $X_i$, $i = 1, \ldots, m$, be normal random variables with means $\mu_i$ and standard deviations $\sigma_i \le \sigma$. Suppose that $|\mu_i - \mu| \le \gamma$ for $i = 1, \ldots, m$, and $c > 0$ is a constant. Then*

$$P\left(\max_{1\le i\le m} |X_i - \mu| \ge \gamma + \sqrt{c \ln m}\sigma\right) \le m^{1-c/2}. \tag{39}$$

PROOF. We shall assume that $m \ge 2$ and that $c \ge 2$, since otherwise the bound is trivial. Denote by $Z$ a standard Gaussian random variable. Then

$$\begin{aligned} P\left(\max_{1\le i\le m} |X_i - \mu| \ge \gamma + \sqrt{c \ln m}\sigma\right) &\le \sum_{i=1}^{m} P(|X_i - \mu| \ge \gamma + \sqrt{c \ln m}\sigma) \\ &\le m P(|Z| \ge \sqrt{c \ln m}\,) \\ &\le m^{1-c/2}. \end{aligned}$$

The last inequality follows from standard bounds on tail probabilities of Gaussian distributions once we note that $c \ln m \ge 1$ when $c \ge 2$ and $m \ge 2$. □

PROOF OF THEOREM 4. Let $\lambda \ge 1$ and $D_\lambda = 3 + \sqrt{32\lambda \ln 2k^3}$ and we will write $D_p$ when $\lambda = p$. Then it is easy to check from Lemma 2 that if $|E\hat{T}_i - Tf| \ge D_\lambda M\omega(\frac{1}{\sqrt{n}}, \mathcal{F})$ then $P(\hat{i} = i) \le \frac{1}{k^{3\lambda-1}}$.

Let

$$I_1 = \left\{i : |E\hat{T}_i - Tf| \ge D_p M\omega\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right)\right\},$$

$$I_2 = \left\{i : |E\hat{T}_i - Tf| < D_p M\omega\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right)\right\}.$$

We then have

$$\begin{aligned}
&E|\hat{T}^* - Tf|^p \\
&\qquad = \sum_{i \in I_1} E(|\hat{T}_i - Tf|^p \mathbb{1}(\hat{i} = i)) + \sum_{i \in I_2} E(|\hat{T}_i - Tf|^p \mathbb{1}(\hat{i} = i)) \\
&\qquad \le \sum_{i \in I_1} (E(\hat{T}_i - Tf)^{2p})^{1/2} (P(\hat{i} = i))^{1/2} + E\Big(\max_{i \in I_2} |\hat{T}_i - Tf|^p\Big).
\end{aligned} \tag{40}$$

Now note that, if $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$ and $p \ge 1$, then a straightforward calculation shows that

$$(EX^{2p})^{1/2} \le 2^p (|\mu^p| + a_{2p}^{1/2} \sigma^p)$$

where $a_j = E|Z|^j$ for $Z$ a standard Gaussian random variable. If $i \in I_1$, then for some $\lambda \ge p$,

$$|E\hat{T}_i - Tf| = D_\lambda M \omega\Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big),$$

and so for $i \in I_1$ and such a choice of $\lambda \ge p$,

$$(E|\hat{T}_i - Tf|^{2p})^{1/2} (P(\hat{i} = i))^{1/2} \le 2^p M^p (a_{2p}^{1/2} + D_\lambda^p) \omega^p \Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big) \frac{1}{k^{(3\lambda - 1)/2}}.$$

Now note that, if $k \ge 2$,

$$\sup_{\lambda \ge p} \frac{D_\lambda^p}{k^{(3\lambda-1)/2}} = \frac{D_p^p}{k^{(3p-1)/2}} \le \frac{D_p^p}{k},$$

and hence

$$\sum_{i \in I_1} (E|\hat{T}_i - Tf|^{2p})^{1/2} (P(\hat{i} = i))^{1/2} \le 2^p M^p (a_{2p}^{1/2} + D_p^p) \omega^p \Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big). \tag{41}$$

Let $m$ be the cardinality of the set $I_2$. Now note that, if $m \le 1$, then

$$E\Big(\max_{i \in I_2} |\hat{T}_i - Tf|^p\Big) \le B(p) \omega^p \Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big) \tag{42}$$

and the theorem now follows from (40)–(42). If $m \ge 2$, then

$$\begin{aligned}
&E\Big(\max_{i \in I_2} |\hat{T}_i - Tf|^p\Big) \\
&\qquad \le (D_p + \sqrt{3 \ln m})^p M^p \omega^p \Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big) \\
&\qquad\quad + M^p \omega^p \Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big) \\
&\qquad\quad \times \sum_{l=3}^{\infty} \Big\{ (D_p + \sqrt{l \ln m})^p
\end{aligned}$$

$$\begin{aligned}
&\times P\biggl((D_p+\sqrt{(l-1)\ln m}\,)M\omega\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr)\le \max_{i\in I_2}|\hat{T}_i-Tf| \\
&\qquad\qquad \le (D_p+\sqrt{l\ln m}\,)M\omega\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr)\biggr)\biggr\} \\
&\le (D_p+\sqrt{3\ln m}\,)^p M^p\omega^p\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr) \\
&\quad + M^p\omega^p\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr) \\
&\quad\quad \times \sum_{l=3}^{\infty}\biggl\{(D_p+\sqrt{l\ln m}\,)^p \\
&\qquad\qquad \times P\biggl(\max_{i\in I_2}|\hat{T}_i-Tf|\ge (D_p+\sqrt{(l-1)\ln m}\,)M\omega\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr)\biggr)\biggr\}.
\end{aligned}$$

Note that it follows from the definition of $I_2$ and the fact that the variance of $\hat{T}_i$ is bounded by $M^2\omega^2(\frac{1}{\sqrt{n}},\mathcal{F}))$ and Lemma 3 that

$$P\biggl(\max_{i\in I_2}|\hat{T}_i-Tf|\ge (D_p+\sqrt{(l-1)\ln m}\,)M\omega\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr)\biggr)\le m^{-(l-3)/2}. \tag{43}$$

Hence

$$\begin{aligned}
&E\Bigl(\max_{i\in I_2}|\hat{T}_i-Tf|^p\Bigr) \\
&\quad\le \biggl[(D_p+\sqrt{3\ln m}\,)^p+\sum_{l=3}^{\infty}(D_p+\sqrt{l\ln m}\,)^p m^{-(l-3)/2}\biggr]M^p\omega^p\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr) \qquad (44)\\
&\quad\le B(p)(\ln k)^{p/2}M^p\omega^p\biggl(\frac{1}{\sqrt{n}},\mathcal{F}\biggr).
\end{aligned}$$

The theorem now follows on combining (40), (41) and (44). □

## 4. Linear estimators.

**4. Linear estimators.** We now consider the performance of linear procedures. As mentioned in the Introduction, the optimal linear procedure is within a small constant factor of the minimax risk when the parameter space is convex. The following theorem considers the case when the parameter space is nonconvex. Let $\mathcal{F}$ denote a parameter set and let C.Hull($\mathcal{F}$) denote the convex hull of $\mathcal{F}$.

THEOREM 5. *Consider the white noise model (1) or the sequence model (2). The minimax linear risk over a parameter set $\mathcal{F}$ is the same as the minimax linear risk over the convex hull of $\mathcal{F}$, that is,*

$$R_A^*(n;\mathcal{F})=R_A^*(n;\mathrm{C.Hull}(\mathcal{F})).$$

This theorem is a direct consequence of the following result.

THEOREM 6. *Let $\hat{T}$ be a linear estimator of $Tf$ where $T$ is a linear functional. Then for any $\mathcal{F}$*

$$\sup_{f\in\mathcal{F}} E_f(\hat{T}-Tf)^2 = \sup_{f\in \mathrm{C.Hull}(\mathcal{F})} E_f(\hat{T}-Tf)^2. \tag{45}$$

PROOF. Since $\mathcal{F}\subseteq \mathrm{C.Hull}(\mathcal{F})$, it is obvious that

$$\sup_{f\in\mathcal{F}} E(\hat{T}-Tf)^2 \le \sup_{f\in \mathrm{C.Hull}(\mathcal{F})} E(\hat{T}-Tf)^2.$$

Let $f\in \mathrm{C.Hull}(\mathcal{F})$ and $f=\sum_i \lambda_i f_i$ with $f_i\in\mathcal{F}$, $\lambda_i\ge 0$ and $\sum_i\lambda_i=1$. Then the squared bias

$$\begin{aligned}(E_f\hat{T}-Tf)^2 &= \Biggl(\sum_i \lambda_i(E_{f_i}\hat{T}-Tf_i)\Biggr)^2 \le \Biggl(\sum_i \lambda_i|E_{f_i}\hat{T}-Tf_i|\Biggr)^2\\ &\le \max_i |E_{f_i}\hat{T}-Tf_i|^2 \le \sup_{f\in\mathcal{F}}(E_f\hat{T}-Tf)^2.\end{aligned}$$

It then follows from the fact that a linear estimator has constant variance that

$$\sup_{f\in \mathrm{C.Hull}(\mathcal{F})} E(\hat{T}-Tf)^2 \le \sup_{f\in\mathcal{F}} E(\hat{T}-Tf)^2. \qquad \square$$

Note that equation (45) is not necessarily true for nonlinear procedures. The following corollary is a direct consequence of Theorem 5.

COROLLARY 1.

$$R_A^*(n;\mathcal{F}) = \sup_{\varepsilon>0}\omega^2(n,\mathrm{C.Hull}(\mathcal{F}))\frac{1/(4n)}{1/n+\varepsilon^2/4} \tag{46}$$

*and*

$$\frac{1}{8}\omega^2\biggl(\frac{2}{\sqrt{n}},\mathrm{C.Hull}(\mathcal{F})\biggr) \le R_A^*(n;\mathcal{F}) \le \frac{1}{4}\omega^2\biggl(\frac{2}{\sqrt{n}},\mathrm{C.Hull}(\mathcal{F})\biggr). \tag{47}$$

Thus the minimax linear risk is determined by the modulus of continuity over the convex hull of $\mathcal{F}$, not over $\mathcal{F}$ itself. In the case that $\omega(\varepsilon,\mathrm{C.Hull}(\mathcal{F}))\gg \omega(\varepsilon,\mathcal{F})$, linear procedures will perform poorly. Examples which illustrate this point are contained in the next section.

**5. Examples.** In this section we discuss examples where the modulus of continuity over the convex hull of the parameter space is much larger than the modulus of continuity over the parameter space. Since the performance of the optimal linear procedure is determined by the modulus of the convex hull of the parameter space linear procedures perform badly in these cases. On the other hand, the nonlinear procedure introduced in Section 3 is within a constant factor of the minimax risk.

5.1. *Estimating functions at a point.* Suppose we observe the white noise model (1) over the interval $[-\frac{1}{2},\frac{1}{2}]$ and we wish to estimate $Tf = f(0)$.

We recall that a function is $\operatorname{Lip}(\alpha)$ $(0<\alpha\le 1)$ over an interval $[a,b]$ if

$$|f(x)-f(y)| \le |x-y|^{\alpha} \qquad \text{for all } x,y\in[a,b].$$

Let

$$\begin{aligned}\mathcal{F}_1 = \{f : f &\text{ is continuous on } [-\tfrac{1}{2},\tfrac{1}{2}]\\ &\text{with maximum at 0 and } f \text{ is } \operatorname{Lip}(1) \text{ over } [-\tfrac{1}{2},0]\}\end{aligned}$$

and

$$\begin{aligned}\mathcal{F}_2 = \{f : f &\text{ is continuous on } [-\tfrac{1}{2},\tfrac{1}{2}]\\ &\text{with maximum at 0 and } f \text{ is } \operatorname{Lip}(\tfrac{1}{2}) \text{ over } [0,\tfrac{1}{2}]\}.\end{aligned}$$

Let $\mathcal{F}=\mathcal{F}_1\cup\mathcal{F}_2$. The parameter spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ are both convex, but $\mathcal{F}$ is nonconvex. It is easy to see that

$$\text{C.Hull}(\mathcal{F}) = \{\text{All continuous functions over } [-\tfrac{1}{2},\tfrac{1}{2}] \text{ with maximum at } 0\}.$$

The convex hull of $\mathcal{F}$ is "much larger" than $\mathcal{F}$. By straightforward calculations it is easy to verify that for $Tf = f(0)$ and small $\varepsilon>0$,

$$\begin{aligned}\omega(\varepsilon,\mathcal{F}_1) &= \omega(\varepsilon,\mathcal{F}_2,\mathcal{F}_1) = 3^{1/3}\varepsilon^{2/3},\\ \omega(\varepsilon,\mathcal{F}_2) &= \omega(\varepsilon,\mathcal{F}_1,\mathcal{F}_2) = 2^{1/4}\varepsilon^{1/2}(1+o(1))\end{aligned}$$

so $\omega(\varepsilon,\mathcal{F}) = 2^{1/4}\varepsilon^{1/2}(1+o(1))$. But $\omega(\varepsilon,\text{C.Hull}(\mathcal{F}))=\infty$.

It follows from Theorem 4 that the minimax mean squared error rate of convergence for estimating the linear functional $Tf=f(0)$ is $n^{-1/2}$. However, the maximum risk of any linear estimator over $\mathcal{F}$ is not even bounded. [This follows from the fact that $\omega(\varepsilon,\text{C.Hull}(\mathcal{F}))=\infty$.] In other words, linear estimators do not work at all in this case.

5.2. *Estimating a linear functional of nearly black objects.* In this example we consider the Gaussian sequence model

$$y_i = f_i + n^{-1/2}z_i, \qquad i=1,\dots,n, \tag{48}$$

where $z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. The size of the vector, $n$, is assumed large; we are interested in asymptotics in which the number of variables is large. We assume that the vector $f$ is sparse: only a small fraction of components are nonzero, and the indices, or locations of the nonzero components are not known in advance.

Denote the $\ell_0$ quasi-norm by $\|f\|_0 = \text{Card}(\{i : f_i \neq 0\})$. Fix $k_n$, the collection of vectors with at most $k_n$ nonzero entries is

$$\mathcal{F} = \ell_0(k_n) = \{f \in \mathbb{R}^n : \|f\|_0 \leq k_n\}.$$

Following Donoho, Johnstone, Hoch and Stern (1992), we call a setting nearly black when the fraction of nonzero components $k_n/n \approx 0$, by analogy with night-sky images. In this example we assume that $k_n$ is known and $k_n \leq Cn^{\varepsilon}$ where $\varepsilon < 1/2$.

A motivation for this model is provided by wavelet analysis, since the wavelet representation of many smooth and piecewise smooth signals is sparse and nearly black in this sense [see, e.g., Donoho, Johnstone, Kerkyacharian and Picard (1995)]. For estimating the whole object, this model has also been studied in Donoho, Johnstone, Hoch and Stern (1992) and Abramovich, Benjamini, Donoho and Johnstone (2000).

In the present paper we are interested in estimating the linear functional of the unknown vector $f$ given by

$$Tf = \sum_{i=1}^{n} f_i.$$

Let $\mathcal{I}(k,n)$ be the class of all subsets of $\{1,\dots,n\}$ of $k$ elements and for $I \in \mathcal{I}(k,n)$ let

$$\mathcal{F}_I = \{f \in \mathbb{R}^n : f_j = 0 \ \forall j \notin I\}.$$

Note that $\mathcal{F}_I$ is a $k_n$-dimensional subspace spanned by the coordinates in $I$. These are obviously convex and $\mathcal{F} = \cup \mathcal{F}_I$ where the union is taken over $I$ in the set $\mathcal{I}(k_n,n)$. From now on we shall assume that $I$ is in the set $\mathcal{I}(k_n,n)$.

Linear procedures perform poorly over $\mathcal{F}$. In fact it is easy to see that the convex hull of $\mathcal{F}$ is the whole of $\mathbb{R}^n$ and

$$\omega(\varepsilon, \text{C.Hull}(\mathcal{F})) = \omega(\varepsilon, \mathbb{R}^n) = \sqrt{n}\varepsilon.$$

It then follows from Theorem 5 that any linear estimator must have maximum mean squared error over $\mathcal{F}$ of at least 1. In fact it is easy to see that the best linear procedure is simply $\hat{T} = \sum_{i=1}^{n} y_i$.

Nonlinear procedures can perform much better. Our general construction given in Section 3 starts with linear estimators constructed assuming that $f \in \mathcal{F}_I$. In this case it is natural to start with $\hat{T}_I$ the minimax estimator over $\mathcal{F}_I$ since this estimator is linear, unbiased over $\mathcal{F}_I$ and has variance equal to $\frac{k_n}{n}$. $\hat{T}_I$ is in fact just the sum of $y_j$ with $j \in I$. In this example equation (29) holds for all $I \in \mathcal{I}(k_n, n)$ with $M = 1$ and $\omega^2(\frac{1}{\sqrt{n}}, \mathcal{F}_I) = \frac{k_n}{n}$.

The construction of $\hat{T}^*$ is also based on the modulus $\omega(\varepsilon, \mathcal{F}_I, \mathcal{F}_J)$ between $\mathcal{F}_I$ and $\mathcal{F}_J$. Note that a least favorable pair of parameters is given by one parameter which has the $k_n$ coefficients in $J$ all equal to some given value $a > 0$ and the rest zero and the second parameter has the coefficients in $J \setminus I$ equal to $-a$ and the rest zero. By choosing $a$ so that the $l_2$ distance between these parameters is equal to $\varepsilon$ it is easy to check that

$$\omega(\varepsilon, \mathcal{F}_I, \mathcal{F}_J) = \sqrt{\operatorname{Card}(I \cup J)}\varepsilon$$

and consequently

$$\omega(\varepsilon, \mathcal{F}) = \sqrt{2k_n}\varepsilon.$$

Now let $\hat{T}_{I,J}$ be defined as in Section 3. It is easy to see in this case that

$$\hat{T}_{I,J} = \sum_{l \in I \cup J} y_l.$$

Let $N$ be the number of parameter spaces. Then $N$ is equal to $n$ choose $k_n$ and it is easy to see that

$$N = \binom{n}{k_n} \le n^{k_n}.$$

It then follows from Theorem 4 that if $\hat{T}^*$ is defined by (33), then

$$\sup_{f \in \mathcal{F}} E(\hat{T}^* - Tf)^2 \le C \frac{k_n^2 \ln n}{n}. \tag{49}$$

The following theorem shows that the estimator $\hat{T}^*$ is in fact rate optimal. The theorem gives a minimax lower bound based on using a mixture prior and a constrained risk inequality introduced in Brown and Low (1996).

Theorem 7. *Let $Tf = \sum_{i=1}^n f_i$. Suppose that $n \ge 4$ and that $k_n < n^\varepsilon$ with $\varepsilon < 1/2$. Then*

$$\inf_{\hat{T}} \sup_{f \in \mathcal{F}} E(\hat{T} - Tf)^2 \ge \frac{1}{121} \frac{k_n^2}{n} \ln\left(\frac{n}{k_n^2}\right). \tag{50}$$

Remark. Comparing the minimax lower bound (50) with the risk upper bound for $\hat{T}^*$, for $k_n < Cn^\varepsilon$ with $\varepsilon < 1/2$, the estimator $\hat{T}^*$ is within a constant factor of the minimax risk. For example, for $k_n = n^\varepsilon$ with $\varepsilon < 1/2$, the risk of $\hat{T}^*$ converges at the rate of $n^{-(1-2\varepsilon)} \log n$ which is optimal.

Proof of Theorem 7. In the proof we will omit the subscript in $k_n$ and simply write $k$ for $k_n$. Let $\psi_\mu$ be the density of a normal distribution with mean $\mu$ and variance $\frac{1}{n}$. And for $I \in \mathcal{I}(k,n)$ let

$$g_I(y_1,\ldots,y_n) = \prod_{j=1}^n \psi_{f_j}(y_j)$$

where $f_j = \frac{\rho}{\sqrt{n}} \mathbb{1}(j \in I)$. Finally let

$$g = \frac{1}{\binom{n}{k}} \sum_{I \in \mathcal{I}(k,n)} g_I$$

and $f = \prod_{j=1}^n f_0$ be the density of $n$ independent normal random variables each with mean 0 and variance $\frac{1}{n}$. Note that a similar mixture prior was used in Baraud (2000) to give lower bounds in a nonparametric testing problem. Now note that if

$$E_{g_I}\left(\delta - k\frac{\rho}{\sqrt{n}}\right)^2 \le C$$

for all $I \in \mathcal{I}(k,n)$ then it follows that

$$E_g\left(\delta - k\frac{\rho}{\sqrt{n}}\right)^2 \le C.$$

We will now apply the constrained risk inequality of Brown and Low (1996). First we need to calculate a chi-squared distance between $f$ and $g$. This is done as follows. Note that

$$\int \frac{g^2}{f} = \frac{1}{\binom{n}{k}^2} \sum_{I \in \mathcal{I}(k,n)} \sum_{I' \in \mathcal{I}(k,n)} \int \frac{g_I g_{I'}}{f}$$

and simple calculations show that

$$\int \frac{g_I g_{I'}}{f} = \exp(j\rho^2)$$

where $j$ is the number of points in the set $I \cap I'$. It follows that

$$\int \frac{g^2}{f} = E \exp(J\rho^2)$$

where $J$ has a hypergeometric distribution

$$P(J = j) = \frac{\binom{k}{j}\binom{n-k}{k-j}}{\binom{n}{k}}.$$

Now note that from Feller [(1968), page 59],

$$P(J=j) \le \binom{k}{j}\left(\frac{k}{n}\right)^j \left(1-\frac{k}{n}\right)^{k-j} \left(1-\frac{k}{n}\right)^{-k}.$$

Now suppose that $n \ge 4$ and that $k < n^{1/2}$. Then

$$\left(1-\frac{k}{n}\right)^{-k} \le 4^{k^2/n} \le 4$$

and hence

$$P(J=j) \le 4\binom{k}{j}\left(\frac{k}{n}\right)^j \left(1-\frac{k}{n}\right)^{k-j}.$$

It now follows that if $n \ge 4$ and $k < n^{1/2}$ then

$$\begin{aligned}\int \frac{g^2}{f} &= E\exp(J\rho^2)\\ &\le 4\left(1-\frac{k}{n}+\frac{k}{n}e^{\rho^2}\right)^k.\end{aligned}$$

Now take $\rho = \sqrt{\ln \frac{n}{k^2}}$ and it follows that

$$\begin{aligned}\int \frac{g^2}{f} &\le 4\left(1+\frac{1}{k}\right)^k\\ &\le 4e.\end{aligned}$$

It then follows from the constrained risk inequality in Brown and Low (1996) that if

$$E_f(\delta-0)^2 \le c_1 \frac{k^2}{n}\ln\frac{n}{k^2} \tag{51}$$

then

$$\begin{aligned}E_g\left(\delta - k\frac{\rho}{\sqrt{n}}\right)^2 &\ge \frac{k^2}{n}\ln\frac{n}{k^2} - 4e\frac{k}{\sqrt{n}}\sqrt{\ln\frac{n}{k^2}}\sqrt{c_1\frac{k^2}{n}\ln\frac{n}{k^2}}\\ &= \frac{k^2}{n}\ln\frac{n}{k^2}(1-4e\sqrt{c_1}).\end{aligned} \tag{52}$$

The theorem now follows on taking $c_1 = 1 + 8e^2 - 4e\sqrt{1+4e^2}$. □

5.3. *Structured nearly black objects.* We will now consider an example under the Gaussian sequence model (48) where most of the coordinates are zero but where we shall also assume that the $k_n$ nonzero coordinates appear consecutively and that $0 \le k_n \le n^{\varepsilon}$ for some $\varepsilon < 1$. Again $k_n$ is assumed to be known. Let

$$\mathcal{F}(a,k_n) = \{f \in R^n : f_i = 0 \text{ unless } a \le i \le a+k_n-1\}$$

and

$$\mathcal{F} = \bigcup_{a=1}^{n-k_n} \mathcal{F}(a, k_n).$$

We call members of $\mathcal{F}$ structured nearly black objects. It is easy to see that the convex hull of $\mathcal{F}$ is again the whole of $\mathbb{R}^n$. It thus follows from Theorem 5 that linear procedures perform poorly for estimating $Tf$ over $\mathcal{F}$.

Let $\hat{T}_a = \sum_{i=a}^{a+k_n-1} y_i$. Then $\hat{T}_{a,b}$ as defined in Section 3 is given by

$$\hat{T}_{a,b} = \sum y_i \mathbb{1}(i \in [a, a+k_n-1] \cup [b, b+k_n-1]).$$

Note that $\mathcal{F}$ is a union of only $n - k_n$ convex sets and so it then follows from Theorem 4 that if $\hat{T}^*$ is now defined by (33) then

$$\sup_{f\in\mathcal{F}} E(\hat{T}^* - Tf)^2 \le C\frac{k_n \ln n}{n}. \tag{53}$$

Equation (53) gives an upper bound for the minimax risk. We shall now show that this upper bound is rate sharp. In fact we shall show that if $n \ge 4$ and $k_n < n^{\varepsilon}$ with $\varepsilon < 1$, then

$$\inf_{\hat{T}} \sup_{f\in\mathcal{F}} E(\hat{T} - Tf)^2 \ge \frac{1}{18}\frac{k_n}{n}\ln\left(\frac{3n}{k_n}\right). \tag{54}$$

This can be seen as follows. Denote the index sets $I_a = \{i : a \le i \le a + k_n - 1\}$ and let $\mathcal{I}(k_n, n) = \bigcup_{a=1}^{n-k_n} I_a$. As in the previous example let $\psi_f$ be the density of a normal distribution with mean $f$ and variance $\frac{1}{n}$. And for $I \in \mathcal{I}(k_n, n)$ let

$$g_I(y_1, \ldots, y_n) = \prod_{j=1}^{n} \psi_{f_j}(y_j)$$

where $f_j = \frac{\rho}{\sqrt{n}}\mathbb{1}(j \in I)$. Finally let $g = \frac{1}{n-k_n}\sum_{I=1}^{n-k_n} g_I$ and $f = \prod_{j=1}^{n} f_0$ be the density of $n$ independent normal random variables each with mean 0 and variance $\frac{1}{n}$. Following the argument in the previous example we note that

$$\int \frac{g^2}{f} = E\exp(J\rho^2)$$

where this time $J$ satisfies

$$P(J = 0) = \frac{n - k_n}{n}$$

and for $1 \le i \le k_n$,

$$P(J = i) = \frac{1}{n}.$$

Hence

$$\int \frac{g^2}{f} \leq 1 + \frac{k_n}{n} \exp(k_n \rho^2).$$

Now set

$$\rho = \sqrt{\frac{1}{k_n}} \sqrt{\ln\left(\frac{3n}{k_n}\right)}.$$

Then $\int \frac{g^2}{f} \leq 4$ and (54) now follows as in (51) and (52).

**Acknowledgment.** We thank two referees for their thorough and useful comments which have helped to improve the presentation of the paper.

## REFERENCES

[1] Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical report 2000-19, Dept. Statistics, Stanford Univ.

[2] Baraud, Y. (2000). Nonasymptotic minimax rates of testing in signal detection. Technical report, Ecole Normale Supérieure.

[3] Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965

[4] Cai, T. and Low, M. (2002). On modulus of continuity and adaptability in nonparametric functional estimation. Technical report, Dept. Statistics, Univ. Pennsylvania.

[5] Donoho, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082

[6] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 41–81. MR1157714

[7] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369. MR1323344

[8] Donoho, D. L. and Liu, R. C. (1991a). Geometrizing rates of convergence. II. *Ann. Statist.* **19** 633–667. MR1105839

[9] Donoho, D. L. and Liu, R. G. (1991b). Geometrizing rates of convergence. III. *Ann. Statist.* **19** 668–701. MR1105839

[10] Efromovich, S. and Low, M. G. (1994). Adaptive estimates of linear functionals. *Probab. Theory Related Fields* **98** 261–275. MR1258989

[11] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York. MR228020

[12] Ibragimov, I. A. and Has'minskii, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32. MR739497

[13] Lepski, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466. MR1091202

[14] Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *Ann. Statist.* **23** 824–835. MR1345202

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania 19104-6340
USA
e-mail: tcai@wharton.upenn.edu
e-mail: lowm@wharton.upenn.edu